CoE: A Clue of Emotion Framework for Emotion Recognition in Conversations

Anonymous ACL submission

Abstract

Emotion Recognition in Conversations (ERC) is crucial for machines to understand dynamic human emotions. While large language models (LLMs) show promise, their performance is often limited by challenges in interpreting complex conversational streams. We introduce a Clue of Emotion (CoE) framework, which progressively integrates key conversational clues to enhance ERC. Building on CoE, we propose a multi-stage auxiliary learning strategy that incorporates tasks like role-playing, speaker identification, and emotion reasoning, improving the model's grasp of conversational and emotional behavior. These tasks strengthen different dimensions of conversational and emotional behavior, deepening the model's understanding of various clues for ERC. Our experiments on EmoryNLP, MELD, and IEMOCAP demonstrate that CoE consistently outperforms state-of-the-art methods, achieving a 2.47% improvement on EmoryNLP. These results underscore the effectiveness of clues and multi-stage auxiliary learning for ERC, offering valuable insights for future research.

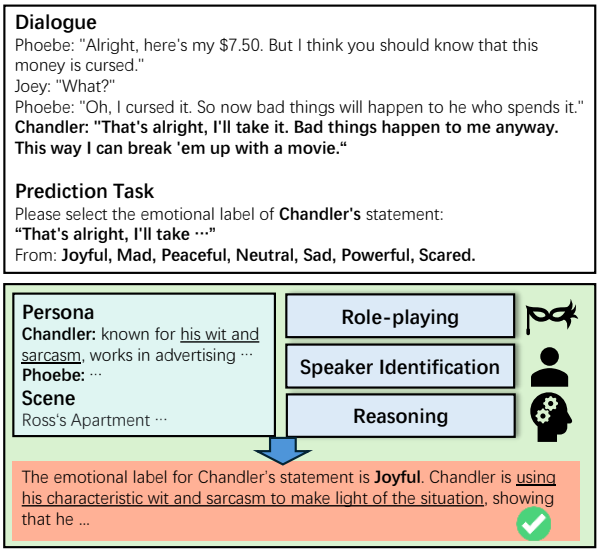

Figure 1: This figure illustrates the CoE framework in ERC task, where contextual information (e.g., Persona and Scene) on the left and auxiliary tasks (e.g., role-playing, speaker identification, and reasoning) on the right are integrated to infer Chandler's emotional label. These combined components enhance the model's ability to reason through emotional clues and determine that Chandler's emotional state is Joyful.

1 Introduction

Emotion Recognition in Conversations (ERC) (Poria et al., 2019; Zahiri and Choi, 2018) has emerged as a significant research direction, driven by its potential applications to improve human-computer interactions (Brave and Nass, 2007), virtual assistants (Chatterjee et al., 2021), and mental health monitoring (Monteith et al., 2022). Accurately identifying and interpreting emotions within conversational contexts is challenging due to the complexity and subtlety of emotional expressions.

Previous works have used recurrent-based models (Hu et al., 2023) and transformer-based discriminative models (Qin et al., 2023) to learn contextual representations, with some methods further employing Graph Neural Networks to model interactions and influences between individuals in conversations (Ghosal et al., 2019). Recently, generative language models (Brown, 2020) have demonstrated powerful zero-shot capabilities, introducing a new paradigm to various natural language processing tasks. In ERC, some approaches have transformed classification tasks into generation tasks by generating class tokens. InstructERC (Lei et al., 2023) is the first to construct an ERC method based on LLMs, utilizing similarity-based retrieval to build context-learning instructions and incorporating multi-task learning to fine-tune the model.

However, previous LLM-based methods (Lei et al., 2023) have overlooked the crucial role of dialogue-specific context and history, limiting their

ability to extract deeper insights from conversations and leading to an incomplete understanding of the emotions conveyed. This highlights the need for a model that can effectively capture and analyze the nuances of dialogue context and character information to provide a more comprehensive understanding of emotions within dialogues.

We reckon that previous research on ERC tasks still needs more exploration of global dialogue scene information, which is a crucial component affecting emotion analysis in conversations (Poria et al., 2017). In light of these considerations, we propose the Clue of Emotion (CoE) framework to deeply explore the diverse textual clues of characters and scenes within dialogues. These clues help the model gain a deep understanding of the emotions in the dialogue. As shown in Figure 1, we incorporate persona and scene as global textual clues to enrich the context of the dialogue and strengthen the model's understanding of conversational dynamics. Additionally, we introduce auxiliary tasks such as role-playing, speaker identification, and reasoning. The first two tasks serve as complementary role-enhancement tasks within the dialogue, enabling the LLM to understand character personalities and scene characteristics deeply. The latter task employs a self-taught emotion rationale reasoning and explanation approach, allowing the model to understand characters' emotions in specific scenarios. Through emotion rationale reasoning, we enhance the language model's ability to infer emotions accurately and logically in complex scenarios. We also design a multi-stage auxiliary learning training strategy for auxiliary learning tasks that effectively guides the LLM to utilize clues better.

The main contributions of this study are summarized as follows:

- We propose the CoE framework, which simultaneously introduces persona and scene-based textual augmentations for dialogue scenarios, and enhances large language models through textual and parametric optimizations.
- We meticulously design three auxiliary tasks within the framework to train the LLM, enabling it to deeply understand and extract character traits and the underlying reasons for emotions in dialogue scenarios.
- We conduct extensive experiments on the entire framework, and the outstanding results demonstrate its effectiveness. Additionally, we carry out in-depth experiments on multiple auxiliary tasks and ERC training strategies to explore the optimal training path for ERC.

2 Preliminaries

In the ERC task, our goal is to identify the speaker's emotion label at each turn of the conversation. Given a set of speakers S and an emotion label set $\mathcal{E}$, each conversation C is represented as $\{(s_1, u_1), (s_2, u_2), \ldots, (s_N, u_N)\}$, where $s_i \in S$ is the speaker and u_i is the utterance of the i-th turn. The model can only utilize previous dialogue utterance $\{(s_1, u_1), (s_2, u_2), \ldots, (s_t, u_t)\}$ as input to predict the emotion label y_t for the t-th utterance. The emotion labels are from a predefined set $\mathcal{E} = \{e_1, e_2, \ldots, e_o\}$, where o is the number of emotion categories.

3 Clue of Emotion

In this section, we first introduce the overall CoE framework (Section 3.1). Next, we present the auxiliary tasks (Section 3.2), encompassing three distinct domains of emotion-related competency. Finally, we discuss the design of training strategy (Section 3.3).

3.1 CoE Framework

To effectively utilize critical conversational clues in the ERC task, we propose a CoE framework that synthesizes essential dialogue components. Our approach not only analyzes the utterance itself, but also integrates the dialogue scenes, speaker personas, and dialogue history, providing a more comprehensive and structured understanding of the conversation. By combining these elements, the CoE framework can be applied in various ERC-related tasks, and can also be utilized in auxiliary tasks to enhance the model's acquisition of emotion-related domain-specific skills.

To demonstrate the advantages of the CoE framework, consider a conversational scenario where multiple participants with distinct personas engage in a discussion. In such interactions, capturing emotional dynamics requires understanding not only the utterances but also the underlying speaker personas, the evolving dialogue context, and the scene. Each speaker may express emotions differently based on their personas, their role in the conversation, and the environment they are situated in. The CoE framework effectively captures subtle

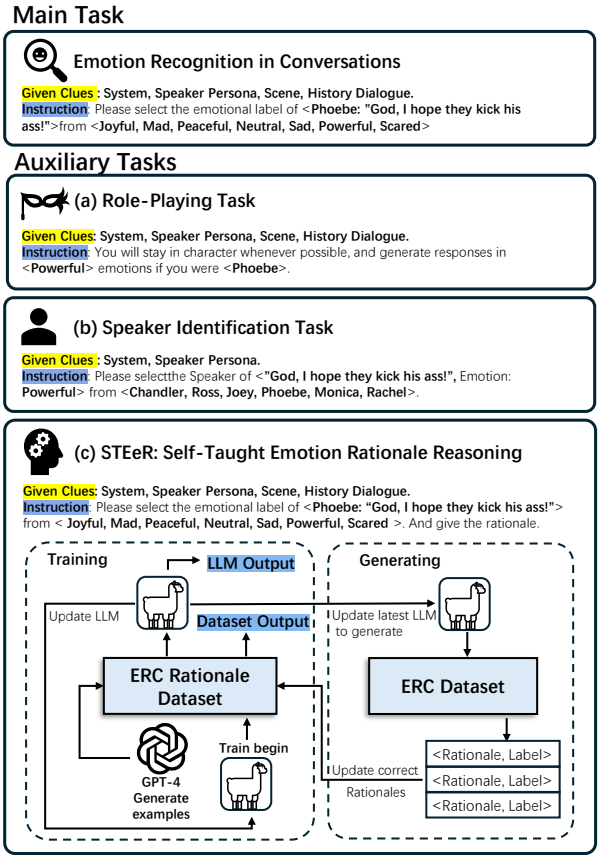

Figure 2: The CoE Framework

emotional shifts by incorporating these key conversational clues. Detailed descriptions of clues can be found in section 4.1.1 the appendix A settings.

To formalize this framework, we define the following components: the speaker set $S = \{s_1, s_2, \ldots, s_M\}$, where M represents the number of speakers; the dialogue scene D; the speaker persona set $P = \{p_1, p_2, \ldots, p_M\}$, where each p_i corresponds to a speaker in S; and the historical context set $H = \{h_1, h_2, \ldots, h_{i-1}\}$, capturing previous dialogue history.

The emotion prediction task using CoE can be expressed as a function f_{emotion}, which combines all these elements to predict the emotion label:

$$e_i = \arg\max_{e \in \mathcal{E}} f_{\text{emotion}}(u_i, S, D, P, H, \mathcal{E}) \quad (1)$$

where e_i is the predicted emotion label for the utterance u_i. Details of the CoE can be seen in Appendix A.

3.2 Auxiliary Learning Tasks

Previous work (Chen et al., 2024; Kung et al., 2021) has shown that identifying and utilizing interdependent skills can improve model performance by making the learning process more efficient. Similarly, we integrate auxiliary tasks related to ERC into the CoE framework to enhance ERC performance by providing additional insights into speaker behavior and conversational dynamics. We focus on three tasks: the role-playing task, the speaker identification task, and the self-taught emotion rationale reasoning (STEeR) task. These tasks capture key conversational elements, helping the model better understand the emotion shift in dialogue. By leveraging these auxiliary tasks, the model can enrich its understanding and boost performance in the ERC task (as shown in Appendix A for details).

3.2.1 Role-Playing Task

Role-specific dialogue generation has been explored in personalizing dialogue agents (Zhang et al., 2018; Madotto et al., 2019), where conditioning on speaker profiles improves response prediction and dialogue coherence by adapting to speaker identity and style. More recently, LLMs have been trained to simulate specific personas by integrating profiles and emotional states (Shao et al., 2023; Wang et al., 2023), further enhancing the quality of generated dialogues. Similarly, role prediction is critical for ERC, as roles shape emotional expression, language style, and intent. The role-playing task enhances the model's ability to generate plausible dialogue by understanding how context, speaker identity, and emotional state influence communication. This ability to predict roles and infer intent is crucial for improving ERC performance within the CoE framework.

In this study, we simulate a role-playing task where the model generates the next utterance u^*_{i+1} based on the current utterance u_i and contextual clues. This task can be formulated as:

$$u^*_{i+1} = f_{\text{role}}(s, e, D, P, H) \quad (2)$$

where s represents the specified speaker, e denotes the speaker's emotion when speaking. The function $f_{\text{role}}(s, e, D, P, H)$ generates the most logically coherent next utterance u^*_{i+1} based on the context H.

To process dialogue segments from various scenarios, we apply a series of filtering and transformation steps. The detailed process is provided in Appendix B.

3.2.2 Speaker Identification Task

Different speakers express emotions in distinct ways. Learning each speaker's characteristics helps

the model understand the link between personal traits and emotional expression. Previous models, like those using hierarchical attention networks and GRUs (Liu et al., 2024), capture emotional nuances but fail to connect these directly to speaker personas, making speaker identification harder. While InstructERC (Lei et al., 2023) introduced speaker identification, it struggled due to missing persona information.

Our CoE framework overcomes this by incorporating speaker personas as key conversational elements. In addition, in our framework, the speaker identification task works complementary to the role-playing task. By jointly learning both tasks, the model benefits from a deeper understanding of how character roles and personalities influence emotional expression. The role-playing task enhances the speaker identification task by providing context on how a speaker's role within the dialogue shapes their emotional communication, leading to better predictions of both identity and intent.

The speaker identification task aims to predict the specific speaker based on the sentence and the emotion expressed by the speaker at the current utterance. The task can be formulated as follows:

$$s^* = \arg\max_{s \in S} f_{\text{speaker}}(u, e, S, P) \tag{3}$$

where u represents the current utterance, e denotes the emotion of the speaker when speaking, and $f_{\text{speaker}}(u, e, S, P)$ is an evaluation function.

3.2.3 Self-Taught Emotion Rationale Reasoning Task

When dealing with limited data samples, the model's self-learning mechanism can enhance performance. Inspired by the Self-Taught Reasoning (STaR) method (Zelikman et al., 2022), we adapted this approach for emotion rationale reasoning, resulting in the development of the **S**elf-**T**aught **E**motion Rational**e** **R**easoning (STEeR) task. This method iteratively generates rationale datasets that improve the model's ability to reason through emotional contexts.

Specifically, STEeR begins by generating a small number of initial rationales using GPT-4[1] due to the large training dataset and GPT-4's relatively low accuracy, which results in a higher proportion of incorrect answers. The model then uses these initial rationales to generate further rationales in subsequent iterations, with feedback guiding the process. Over time, as the model generates more correct rationales, its reasoning accuracy improves progressively through each iteration. This iterative process strengthens the model's emotion reasoning capabilities by continuously refining its ability to generate accurate rationales. This is visually represented in Figure 2, where the rationale generation and feedback loop are depicted.

[1]We use the GPT-4-0125-preview version.

Algorithm 1: Main STEeR Training Process

Input: Pretrained LLM M, Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{D}$, Training Dataset $\mathcal{D}_{\text{train}}$

$M_0 \leftarrow M$, $D_0 \leftarrow D \setminus \mathcal{D}_{\text{train}}$ `// Copy the original model and reasoning dataset (pre-generated)`
for $n \in 1 \ldots N$ **do**
  $M_n \leftarrow \text{train}(M_{n-1}, \mathcal{D}_{\text{train}})$ `// Finetune the model on the updated memory dataset`
  for $(x_i, y_i) \in \mathcal{D}_{n-1}$ **do**
    $(\hat{r}_i, \hat{y}_i) \leftarrow M_n(x_i)$ `// Perform rationale generation`
    $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \{(x_i, \hat{r}_i, y_i) \mid \hat{y}_i = y_i\}$ `// Add correct rationales`
  end
  $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \setminus \mathcal{D}_{\text{train}}$ `// Update dataset by removing generated samples`
end

The STEeR method operates in two stages: the initial data generation phase and the main STEeR training phase (see Algorithm 1). In the training phase, the model is fine-tuned iteratively using the generated rationales, with each cycle refining its emotional reasoning capacity. This integration of rationale generation and model fine-tuning allows for continuous improvement in emotion reasoning tasks, in alignment with the objectives of the CoE framework. The data generation process is detailed in Appendix 2.

3.3 Designing Multi-Stage Training Strategies

Each auxiliary task plays a key role in improving the performance of the ERC. Role-playing helps the model predict character roles, interpret language styles, and infer speaker intent, while speaker identification allows the model to differ-

entiate speakers and grasp their unique emotional expressions. STEeR further develops the model's ability to reason about emotions, ensuring it can generate and justify coherent emotional responses. These tasks collectively form the foundation of a robust ERC model.

Determining the optimal training strategy for these tasks is challenging. A curriculum learning approach (Soviany et al., 2022; Wang et al., 2021), where tasks are trained sequentially, may allow the model to gradually build up essential skills. Alternatively, a mixed training strategy (Chen et al., 2024, 2022), where tasks are trained concurrently, could foster better knowledge transfer and generalization across tasks.

We explored both approaches (see Section 4.4 for results) and found that a multi-stage, mixed auxiliary training strategy delivered the most significant improvements in ERC performance.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We conduct experiments on three well-known benchmarks: MELD (Poria et al., 2019), EmoryNLP (Zahiri and Choi, 2018), and IEMOCAP (Busso et al., 2008).

MELD (Poria et al., 2019) This dataset is derived from the Friends TV series, including over 1400 dialogues and 13000 utterances. Each dialogue involves multiple speakers and is annotated with seven emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. Additionally, we manually set persona information based on the six main characters from the Friends TV series to improve dialogue understanding. We also extract the scenes context for each dialogue opening in the dataset.

EmoryNLP (Zahiri and Choi, 2018) This dataset comprising 97 episodes and 12,606 utterances, with each utterance being annotated with seven emotions adapted from the Willcox's feeling wheel (Willcox, 1982): Sad, Mad, Scared, Powerful, Peaceful, Joyful, and Neutral. We also extract scene information to provide clues about the setting of each dialogue. Similar to MELD, we manually set persona information based on the six main characters from the Friends TV series.

IEMOCAP(Busso et al., 2008) This dataset contains video recordings of dyadic interactions, capturing natural emotional exchanges between pairs of actors. It includes a total of 151 sessions, with each session involving two speakers, resulting in 302 recorded dialogues. Due to the dataset being composed of segmented video recordings, it is hard to construct a complete CoE framework as in other datasets. As a result, we rely on historical dialogue information to assist in the ERC task.

4.1.2 Metrics

To maintain consistency with previous methods, we use weighted F1 as the evaluation metric, as it better reflects the model's true performance given the severe class imbalance in the tested datasets (Lei et al., 2023; Song et al., 2022). Furthermore, we ensured consistency in training/validation/test splits with InstructERC (Lei et al., 2023) and COSMIC (Ghosal et al., 2020).

4.1.3 Training Setup

For our experiments, we ensured that the LLM parameter count was kept around 7B. Considering computational constraints and time limitations, we employed the LoRA method from PEFT (Parameter-Efficient Fine-Tuning) for model fine-tuning. The rank of the LoRA adapter was set to 16, the learning rate was configured to 2e-4, and the LoRA scaling factor was set to 0.1. Our experiments were conducted on an 8x32GB Nvidia V100 GPU cluster, using FP16 precision for training. The final results were averaged across three independent runs.

4.1.4 Baselines

We select several ERC baselines to compare with our CoE. SPCL-CL (Song et al., 2022), COSMIC (Ghosal et al., 2020) uses a transformer-based model, and SPCL-CL introduced a contrastive learning approach. UniMSE (Hu et al., 2022) uses multimodal data. SACL (Hu et al., 2023) and EmotionIC (Liu et al., 2024) use recurrent-based method. DualGATs (Zhang et al., 2023) use GNN-based method. InstructERC (Lei et al., 2023) uses LLM-based methods and also evaluates performance using a parameter-free variant of GPT-3.5 (Ouyang et al., 2022) as an additional baseline.

4.2 Main Results

Previous methods, such as COSMIC (Ghosal et al., 2020), leveraged commonsense knowledge through knowledge graphs to model mental states and contextual information, enhancing their ability to infer emotions. InstructERC (Lei et al., 2023) was the first to leverage generative large language models for ERC tasks, leading to improved performance

across all datasets and demonstrating the potential of generation-based modeling for emotion recognition in conversations. While these models incorporated external information sources, our approach directly models conversational elements such as speaker personas and textual clues, embedding them explicitly within the task itself. In addition, we incorporate auxiliary tasks to further strengthen the model's ability to process these clues by using supervised learning to refine its understanding of emotional dynamics.

Table 1: Main Results for ERC

Methods	Backbone	EmoryNLP W-F1	MELD W-F1	IEMOCAP W-F1	Average W-F1
COSMIC	TB	38.11	65.21	65.28	56.2
SPCL-CL	TB	40.94	67.25	69.74	59.31
UniMSE	TB (MM)	-	65.51	70.66	-
EmotionIC	RB	40.25	66.32	69.61	58.73
SACL	RB	39.65	66.45	69.22	58.44
DualGATs	GB	40.69	66.90	67.68	58.42
GPT3.5	LLM	37.00	65.07	53.38	51.82
InstructERC	LLM	41.37	69.15	**71.39**	60.64
CoE	LLM	**44.29**	**70.11**	70.46	**61.62**

Backbone abbreviations: TB = Transformer-based, MM = Multimodal, RB = Recurrent-based, GB = GNN-based, LLM = Large Language Model.
GPT3.5 results were obtained through the InstructERC framework's testing methods.

Our evaluations (Table 1) show that CoE outperformed previous methods across multiple datasets, achieving a 2.92% improvement on the EmoryNLP dataset and a 0.96% improvement on MELD, reflecting its effectiveness in various ERC benchmarks. Additionally, CoE achieved the highest average W-F1 score (+ 0.98%) across the tested datasets, indicating its overall superiority in handling a variety of conversational contexts. These improvements arise from CoE's ability to progressively integrate conversational clues like speaker personas and context, which are embedded directly within the task through the multi-stage auxiliary learning strategy. By enhancing its understanding of emotional subtleties and speaker-specific behaviors, CoE surpasses prior methods that rely on external knowledge sources. However, on the IEMOCAP dataset, which consists of scripted, discrete conversations between temporary actors, our method faced challenges due to the absence of persona and scene information, which are essential components for effectively capturing the full emotional context.

4.3 Ablation Studies

4.3.1 Effect of Different LLMs

To further evaluate the performance of LLMs in ERC tasks, we conducted experiments using different LLMs, as shown in Figure 3. The experiments focused on the ERC task using the Emotion CoE framework without auxiliary tasks, allowing for a direct comparison of various LLMs' performance.

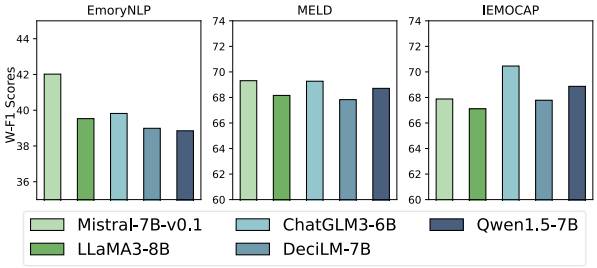

Figure 3: Comparison with Other LLMs

Among the models tested, Mistral-7B-v0.1 demonstrated the best overall performance across the majority of datasets, leading to its selection as the base model for subsequent ablation studies on multi-stage auxiliary learning.

4.3.2 Ablation Study on Textual Clues

In this section, we conduct an ablation study to analyze the impact of textual clues, specifically persona and scene information, on the model's performance within the CoE framework. Importantly, these experiments do not incorporate auxiliary tasks, allowing for a focused analysis of the direct contributions of persona and scene clues. The results are summarized in Table 2. The ablation study highlights the significance of persona and scene clues. Removing either persona or scene leads to a noticeable performance drop, indicating their essential role in providing context to the dialogue. The most significant decline occurs when both clues are removed, underscoring their complementary effect on improving the accuracy of the ERC.

Table 2: Ablation Study on Textual Clues

Methods	EmoryNLP W-F1	MELD W-F1
LLM + Textual Clues	**42.02****	**69.31****
LLM + Textual Clues (w/o Persona)	41.51*	69.13**
LLM + Textual Clues (w/o Scene)	40.71**	-
LLM + Textual Clues (w/o Persona + Scene)	39.95**	-

Note: *p<0.05, **p<0.01 indicate the level of statistical significance.

4.3.3 Ablation Study on Auxiliary Tasks

We conducted an in-depth analysis of the auxiliary tasks to evaluate their individual contributions. Table 3 illustrates the contribution of each auxiliary task to the main task's performance across

both datasets, demonstrating that each task independently adds value to the model's emotion recognition capabilities.

Table 3: Performance Improvement of the Main Task with Each Individual Auxiliary Task

Tasks	EmoryNLP W-F1	MELD W-F1
Emotion CoE Only	42.02**	69.31**
Emotion CoE + Speaker CoE	43.10**	69.67**
Emotion CoE + Role-Playing CoE	42.40*	69.42*
Emotion CoE + Reasoning CoE	43.03**	69.58*

Note: *p<0.05, **p<0.01 indicate the level of statistical significance.

Based on the results, the **speaker identification task** (+1.08% EmoryNLP, +0.36% MELD) demonstrates the strongest effect on model performance, likely because it enables the model to link emotional clues directly to individual personas, capturing subtle shifts in tone that vary between speakers. Although the **role-playing task** (+0.38% EmoryNLP, +0.41% MELD) shows relatively smaller gains, its full potential may be realized when integrated with other auxiliary tasks during joint training, where it can enhance the model's understanding of personas. While **STEeR's** (+1.01% EmoryNLP, +0.27% MELD) improvement is a little smaller than **speaker's**, its focus on emotion rationale reasoning still plays a key role in interpreting complex emotional states. The smaller improvement in MELD (+0.36%) may be due to the imbalanced distribution of emotion classes, where the model focuses more on majorty neutral emotions, leading to limited gains for rarer emotions.

4.4 Train Strategy Study

In this section, we examine various training strategies aimed at enhancing model performance on the ERC task. The experiments were conducted in two phases: first, we performed a preliminary evaluation on a subset of the data to compare the effectiveness of mixed learning and curriculum learning. Building on these findings, we then developed and tested two refined strategies for comprehensive evaluation.

4.4.1 Phase 1: Preliminary Evaluation of Training Approaches

In this initial phase, we conducted experiments on a subset of the dataset to assess the effectiveness of two key training strategies: curriculum-based sequential training (Soviany et al., 2022; Wang et al., 2021) and mixed auxiliary training (Chen et al., 2022). Given the high computational cost associated with training large models across various auxiliary task combinations, we focused on role-playing and speaker identification tasks using the EmoryNLP dataset for evaluation. These tasks were selected because they address critical aspects of the ERC, allowing us to more precisely evaluate the impact of different training strategies.

We compare three approaches, as shown in Figure 4 :

1. **Curriculum-based Sequential Training Strategy:** Each task is trained in a separate stage, with the ERC task always being the final stage of training.

2. **Mixed Auxiliary Training Strategy:** All auxiliary tasks are merged into a single multi-task dataset for joint training, with a separate stage reserved for fine-tuning the model on the ERC task.

3. **Multi-Task Training Strategy:** All tasks, including the ERC task, are mixed and trained together in a single stage as a multi-task setup.

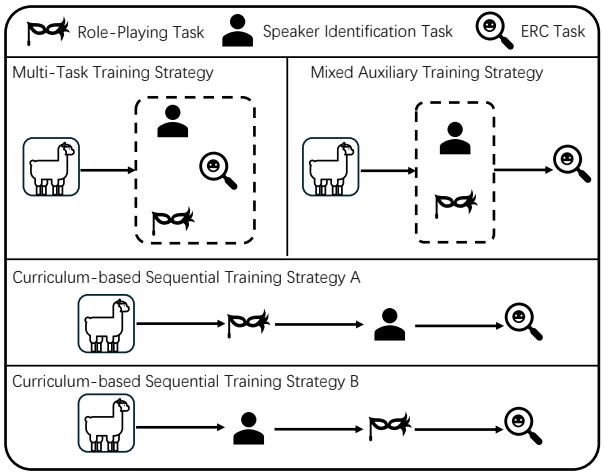

Figure 4: Three Training Strategies for ERC

The results in Table 4 of this small-scale experiment demonstrate that **mixed auxiliary training** consistently outperforms the curriculum-based approach. This suggests that the mixed training strategy allows for more dynamic knowledge sharing between tasks, which makes it a more suitable approach for ERC.

Table 4: Ablation Study On Different Strategies

Models	EmoryNLP W-F1
Fully Mixed Strategy	42.46**
Mixed Auxiliary Strategy	**43.64****
Curriculum-based Strategy A	43.30*
Curriculum-based Strategy B	43.35*

Note: *p<0.05, **p<0.01 indicate the level of statistical significance.

Table 5: Comparison of CoE Strategies A and B with Significance Testing

Strategies	EmoryNLP W-F1	MELD W-F1
CoE Strategy A	43.68*	69.89*
CoE Strategy B	**44.29****	**70.11***

Note: *p<0.05, **p<0.01 indicate the level of statistical significance.

4.4.2 Phase 2: Strategy Refinement and Full Experiment

One of the key challenges identified during our experiments is that the STEeR method cannot be jointly trained with other datasets due to its unique rationale generation process. This limitation means that a straightforward mixed auxiliary training strategy, which typically involves training all tasks together, is not working for integrating STEeR with other tasks. Based on these findings from the small-scale experiments, we developed two refined strategies for full-scale evaluation to address these issues and optimize the training process.

- **Strategy A**: This strategy follows a three-stage training process. First, the model is trained on a combined dataset consisting of role-playing and speaker identification tasks. The model is then trained using the STEeR method. Finally, the model is fine-tuned on the ERC task.

- **Strategy B**: This approach starts by generating a rationale dataset with the STEeR method. This dataset allows for joint training on all three auxiliary tasks. Once the joint training phase is complete, the model is fine-tuned on the ERC task.

Analysis of Strategy Refinement Strategy A follows a structured, stepwise approach, where each training stage builds upon the previous one, progressively enhancing the model's grasp of speaker identities, conversational roles, and emotional reasoning. This hierarchical progression allows the model to specialize in each aspect before moving to the next, creating a refined understanding at every level. On the other hand, Strategy B starts by generating a rationale dataset using the STEeR method, which enables joint training of all three auxiliary tasks simultaneously. This joint approach fosters a more integrated prior model, providing a solid foundation for the ERC task.

We conducted experiments to compare the performance of the model under Strategy A and Strategy B. As shown in Table 5, the results indicate that Strategy B consistently outperformed Strategy A in the final ERC task. By pre-generating a rationale dataset, Strategy B facilitated a more effective integration of auxiliary tasks within its multi-stage auxiliary learning framework, enabling joint training that strengthened the model's overall capability in ERC. In contrast, segmentation in Strategy A can lead to weaker knowledge retention in earlier stages, potentially limiting the model's capacity to fully leverage learned representations in the final ERC stage. These findings underscore the significance of integrating reasoning processes early in the training pipeline, which plays a crucial role in enhancing the model's performance across various ERC benchmarks by providing it with robust prior knowledge.

5 Conclusion

In this work, we present the CoE framework, integrating various textual clues such as personas and scene information to enhance the model's ability to understand emotions within conversations. To capture clues of conversational dynamics and emotional reasoning, we designed three tasks, role-playing, speaker identification, and STEeR. The first two tasks complement each other by improving the model's understanding of textual clues, while STEeR enhances the model's reasoning ability by using clues, leading to more accurate ERC.

Through comprehensive exploration of multi-stage auxiliary training strategies, we discovered that combining auxiliary tasks into a unified multi-task training phase yields better results than the stepwise training approach. Experimental results demonstrate that this strategy improves the model's emotion capability by effectively learning prior knowledge and achieving state-of-the-art performance on the EmoryNLP and MELD datasets.

Limitations

Despite the promising results, our work has limitations. The CoE framework, while effective for text-based ERC, does not incorporate multimodal information such as visual or acoustic cues, which could further enhance emotion recognition. In future work, we aim to explore zero-shot methods like chain-of-thought reasoning and retrieval-augmented generation techniques to improve the model's ability to uncover character-related emotional factors in more complex, multimodal conversations. These methods will help the model better infer emotions in unseen scenarios by incorporating richer contextual data through intermediate reasoning steps or external information retrieval. Additionally, due to computational resource constraints, our experiments were limited to models no larger than 7B parameters.

References

Scott Brave and Cliff Nass. 2007. Emotion in human-computer interaction. In *The human-computer interaction handbook*, pages 103–118. CRC Press.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Carlos Busso, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Rajdeep Chatterjee, Saptarshi Mazumdar, R Simon Sherratt, Rohit Halder, Tanmoy Maitra, and Debasis Giri. 2021. Real-time speech emotion analysis for smart home assistants. *IEEE Transactions on Consumer Electronics*, 67(1):68–76.

Hong Chen, Xin Wang, Chaoyu Guan, Yue Liu, and Wenwu Zhu. 2022. Auxiliary learning with joint task and data scheduling. In *International Conference on Machine Learning*, pages 3634–3647. PMLR.

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. 2024. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, 36.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10835–10852, Toronto, Canada. Association for Computational Linguistics.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Po-Nien Kung, Sheng-Siang Yin, Yi-Cheng Chen, Tse-Hsuan Yang, and Yun-Nung Chen. 2021. Efficient multi-task auxiliary learning: selecting auxiliary data by feature similarity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 416–428.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Yingjian Liu, Jiang Li, Xiaoping Wang, and Zhigang Zeng. 2024. Emotionic: emotional inertia and contagion-driven dependency modeling for emotion recognition in conversation. *Science China Information Sciences*, 67(8):1–17.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Scott Monteith, Tasha Glenn, John Geddes, Peter C Whybrow, and Michael Bauer. 2022. Commercial use of emotion artificial intelligence (ai): implications for psychiatry. *Current Psychiatry Reports*, 24(3):203–211.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Xiangyu Qin, Zhiyu Wu, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, Li Wang, and Jinshi Cui. 2023. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13492–13500.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713*.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

D. Willcox. 1982. The feeling wheel. *Journal of Counseling Psychology*.

Samira M Zahiri and Jinho D Choi. 2018. Emotion recognition in conversations with transfer learning from generative conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 27–36.

Eric Zelikman, Yuhuai Wu, and Noah D Goodman. 2022. Star: Self-taught reasoner. *arXiv preprint arXiv:2203.14465*.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023. Dualgats: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7395–7408.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Details of CoE

In this study, we develop a CoE framework to fine-tune pre-trained language models for emotion-related tasks through supervised fine-tuning (SFT). Each item in the square bracket represents a key element, with the following descriptions:

- **[System] (Predefined Role):** This part outlines the type of task the participant is required to perform (e.g., ERC, role-playing, or speaker identification), as well as the role the large language model needs to assume during the task.
- **[Speaker_Persona] (Character Personality Description):** This part provides detailed descriptions of the characters' personality traits, professional backgrounds, and roles within the group. These descriptions help in establishing a clearer understanding of each character's persona.
- **[Scene] (Setting):** This section describes the specific location and environment where the dialogue takes place, such as "Monica and Rachel's apartment." Setting the scene is crucial for building the context of the conversation.
- **[History_Dialogue] (Historical Dialogue):** This part records prior interactions between the characters, allowing the model to better understand the current conversational context and the dynamics between the participants.
- **[Predict] (Prediction):** This section provides instructions to guide the model in making predictions. For ERC tasks, this might involve predicting the emotion of a specific dialogue segment; for role-playing, it entails predicting what the character will say next; and for speaker identification, it involves predicting the speaker of a particular utterance.
- **[Target]:** This represents the model's expected output, such as the prediction results that the model needs to learn.

B Dialogue Processing Steps

To process dialogue segments from various scenarios, we follow these steps:

1. For each dialogue segment u_i, we remove the first utterance. If u_i still contains more than one utterance, we let the first utterance be $u_i[1]$. The utterance set becomes $u_i \setminus u_i[1]$:

$$F_1(C) = u_i \setminus u_i[1] \tag{4}$$

2. We filter out speakers not in P:

$$F_2(C) = (s, t) \in u_i \mid s \in P \tag{5}$$

where $u_i[1]$ is the first utterance in each dialogue segment C. For each dialogue segment u_i, we retain utterances where the speaker s is in P.

Combining the above steps, the processed dialogue set $P(C)$ is:

$$P(C) = \bigcup_{i=1}^{n} F_2(F_1(u_i)) \tag{6}$$

This indicates that for each dialogue segment u_i in a scenario, we discard the first utterance F_1, then filter out the speakers F_2, and merge all processed segments to form the final dialogue set $P(C)$.

C Initial Data Generation Phase

Algorithm 2: Pre Data Generation Process

Input: GPT Model $GPT4$, Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{D}$

$\mathcal{D}_0 \leftarrow \mathcal{D}$, $\mathcal{D}_{\text{train}} \leftarrow \emptyset$ `// Initialize a temporary dataset and the training dataset`

for $n \in 1 \ldots N$ **do**

  for $(x_i, y_i) \in \mathcal{D}_{n-1}$ **do**

    $(\hat{r}_i, \hat{y}_i) \leftarrow GPT4(x_i, y_i) \quad \forall (x_i, y_i)$ `// Perform rationale generation`

    $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \{(x_i, \hat{r}_i, y_i) \mid \hat{y}_i = y_i\}$ `// Add correct rationales`

  end

  $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \setminus \mathcal{D}_{\text{train}}$ `// Update dataset by removing generated samples`

end

Prompt for ERC	
[System]	As an expert in Conversational Emotion Recognition (ERC), your task is to [Predict] the emotion by analyzing the [Speaker_Persona], [Scene] and [History_dialogue] information.
[Speaker_Persona]	Chandler: Known for wit and sarcasm, works in advertising, group's humorist with logical insights. Phoebe: A masseuse and musician, eccentric with unique wisdom, offering empathy to the group. Ross: A paleontologist and professor, adds intellectual depth, connections through family ties with Monica and past relationship with Rachel. Rachel: Progresses from a waitress to fashion, influences group dynamics, evolving relationship with Ross. Monica: A chef, known for organization, group's anchor with hospitality, Ross's sister.
[Scene]	Monica and Rachel's place, the entire gang is present.
[History_Dialogue]	Monica: "So ah, Phoebe, how was your date?" Phoebe: "Oh well y'know." Monica: "Yeah, I do know." ... Chandler: "Don't worry." Phoebe: "God, I hope they kick his ass!"
[Predict]	Please select the emotional label of <Phoebe: "God, I hope they kick his ass!">from <Joyful, Mad, Peaceful, Neutral, Sad, Powerful, Scared>.
[Target]	Powerful

Prompt for Speaker Recognition Task	
[System]	You are now an expert in Conversational Emotion Recognition. Your task is to predict the speaker by analyzing sentences. Here's a list of individuals and their [Speaker_Persona], which will assist you in understanding them better
[Speaker_Persona]	Chandler: Known for his wit and sarcasm, works in advertising, group's humorist with logical insights. Ross: A paleontologist and professor, adds intellectual depth through family ties with Monica and past relationship with Rachel. Joey: An aspiring actor, the group's innocent and charming member, deepening bond with Chandler through loyalty and simplicity. Phoebe: A masseuse and musician, introduces eccentricity and unique wisdom, offering empathy to the group. Monica: A chef, known for organization, group's anchor with hospitality, directly linking to Ross as his sister. Rachel: Progresses from a waitress to fashion, influencing group dynamics, especially her evolving relationship with Ross.
[Predict]	Please select the Speaker of <"God, I hope they kick his ass!", Emotion: Powerful >from <Chandler, Ross, Joey, Phoebe, Monica, Rachel>.
[Target]	Phoebe

Prompt for Roleplay Task	
[System]	You are now an expert roleplay actor. Analyze the character's [Speaker_Persona], [Scene], and [History_dialogue] to predict their most likely next line.
[Speaker_Persona]	Monica: Known for organization, acts as the group's anchor, fostering cohesion with hospitality, linking to Ross as his sister. Rachel: Progresses from a waitress to a career in fashion, influencing group dynamics, especially her relationship with Ross. Ross: A paleontologist and professor, adds intellectual depth through family ties with Monica and past relationship with Rachel. Chandler: Known for wit and sarcasm, works in advertising, the group's humorist with logical insights. Phoebe: A masseuse and musician, introduces eccentricity and unique wisdom, offering empathy to the group.
[Scene]	Monica and Rachel's, The entire gang is there.
[History_Dialogue]	Monica: "So ah, Phoebe, how was your date?" Phoebe: "Oh well y'know." Monica: "Yeah, I do know." ... Chandler: "Don't worry."
[Predict]	You will stay in character whenever possible, and generate responses in <Powerful>emotion as if you were <Phoebe>.
[Target]	God, I hope they kick his ass!

Prompt for Reasoning Task (STEeR)	
[System]	As an expert in Conversational Emotion Recognition (ERC), your task is to [Predict] the emotion by analyzing the [Speaker_Persona], [Scene] and [History_dialogue] information.
[Speaker_Persona]	Monica: Known for organization, acts as the group's anchor, fostering cohesion with hospitality, linking to Ross as his sister. Rachel: Progresses from a waitress to a career in fashion, influencing group dynamics, especially her relationship with Ross. Ross: A paleontologist and professor, adds intellectual depth through family ties with Monica and past relationship with Rachel. Chandler: Known for wit and sarcasm, works in advertising, the group's humorist with logical insights. Phoebe: A masseuse and musician, introduces eccentricity and unique wisdom, offering empathy to the group.
[Scene]	Monica and Rachel's, The entire gang is there.
[History_Dialogue]	Monica: "So ah, Phoebe, how was your date?" Phoebe: "Oh well y'know." Monica: "Yeah, I do know." ... Chandler: "Don't worry."
[Predict]	Please select the emotional label of <Phoebe: "God, I hope they kick his ass!">from <Joyful, Mad, Peaceful, Neutral, Sad, Powerful, Scared>: And give the rationale.
[Target]	[Rationale] The emotional label of <Phoebe: "God, I hope they kick his ass!">is <Powerful>. In the dialogue, Phoebe starts feeling vulnerable about her date but transitions to confidence after receiving support from her friends. Her assertive statement, indicating a desire for action and control, reflects an empowered emotional state. Thus, "Powerful" is the appropriate label.